

Improved Optimization for the Cluster Jastrow Antisymmetric Geminal Power and Tests on Triple-Bond Dissociations

Eric Neuscamman^{1,2,*}

¹*Department of Chemistry, University of California, Berkeley, California 94720, USA*

²*Chemical Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA*

(Dated: March 23, 2016)

We present a novel specialization of the variational Monte Carlo linear method for the optimization of the recently introduced cluster Jastrow antisymmetric geminal power ansatz, achieving a lower-order polynomial cost scaling than would be possible with a naive application of the linear method and greatly improving optimization performance relative to the previously employed quasi-Newton approach. We test the methodology on highly multi-reference triple-bond stretches, achieving accuracies superior to traditional coupled cluster theory and multi-reference perturbation theory in both the typical example of N_2 and the transition-metal-oxide example of $[ScO]^+$.

I. INTRODUCTION

One of the most pressing problems in quantum chemistry today is the challenge of predicting the detailed effects of electron correlation in systems far from the mean-field regime such as molecules with stretched bonds, transition metal oxide catalysts, and π -conjugated molecules with low-lying doubly-excited states. While traditional quantum chemistry methods that build up from a Hartree-Fock reference function are very effective at describing weak electron correlation (i.e. correlation that does not greatly alter the mean-field picture), and recent advances in density matrix renormalization group (DMRG) [1] and full configuration interaction quantum Monte Carlo (FCI-QMC) [2] have greatly expanded the reach of active space approaches to strong correlation (i.e. correlation, typically within the valence electrons, that causes qualitatively non-mean-field effects), it remains difficult to affordably and accurately describe both weak and strong correlation simultaneously. Recently, we introduced [3] the cluster Jastrow antisymmetric geminal power (CJAGP) ansatz as a candidate to address this challenge by attempting to combine the strengths of cluster operators [4], Hilbert space Jastrow factors [5], and pairing wave functions, but we were limited in our ability to test this new ansatz by the difficulty of combining quasi-Newton optimization techniques with variational Monte Carlo (VMC). In this paper, we present a more robust and efficient optimization scheme for the CJAGP based on the VMC linear method (LM) [6–9] and use it to test this new ansatz on two challenging triple-bond dissociations that were inaccessible to the old optimization method.

The ability of the CJAGP to encode strong correlation arises from its Jastrow-modified geminal power reference [10], and so in a sense the theory can be seen as being part of the chemistry community’s larger effort to construct ansatzes based on electron pairs. In-

deed, the ubiquity of electron pairing in molecular physics has spurred the investigation of numerous pair-based approaches to electron correlation, in which the fundamental wave function building block is a two-electron geminal rather than a one-electron orbital. Early examples include perfect pairing (PP) [11, 12], the “bare” (i.e. not Jastrow-modified) antisymmetric geminal power (AGP) [13–15], and products of strongly orthogonal geminals [16–18]. More recently, there has been renewed interest in pairing wave functions based on the idea of relaxing the strong orthogonality constraint, as in generalizations of PP [19–21] the antisymmetric product of 1-reference-orbital geminals (AP1roG) [22–27] and extensions of the singlet-type strongly orthogonal geminal (SSG) approach [28–33]. While the CJAGP has strong connections to these pairing theories, it is important to recognize that Jastrow-modification can drastically change the ansatz, and it is actually the combination of Jastrow factor and geminal power that lies at the heart of the ansatz’s ability to capture strong correlation [10]. For this reason, the pairing theory that most closely relates to CJAGP is JAGP with real space Jastrows [34–37], although we must emphasize that real space and Hilbert space Jastrow factors are quite different, and so many of the approximations involved are distinct.

The ability of the CJAGP to encode weak correlation arises from the fact that under a unitary orbital rotation, the Hilbert space Jastrow factor becomes a simplified coupled cluster (CC) doubles operator [3] similar in structure to the tensor hypercontraction representation of doubles amplitudes [38]. The variational freedom of the cluster-Jastrow (CJ) operator is much reduced compared to the traditional CC doubles operator [4], and as we will discuss below this simplicity may limit the CJAGP’s ability to encode the finer details of dynamic correlation. Note that the CJ operator is *not* a pairing operator, and that the electron pairing qualities of CJAGP come instead from its AGP reference. One must therefore be careful not to confuse the CJ operator with the CC operator representations of various pairing theories, such as PP [39, 40], some forms of the generalized valence bond [41], AP1roG [23–25], and pair CC

*Electronic mail: eneuscamman@berkeley.edu

doubles [42–44]. Indeed, these theories often use their pairing ansatzes’ cluster operator formulation to facilitate a non-variational, projective optimization scheme as in traditional CC theory, whereas CJAGP is evaluated using *variational* Monte Carlo. As such, it may be conceptually more useful to see CJAGP as an attempt to achieve a type of variational, multi-reference CC, inspired by the accuracy seen in studies of variational and quasi-variational CC [45–50] and the extraordinary accuracies achievable by multi-reference CC [51].

The remainder of this paper is organized as follows. We begin by defining the CJAGP ansatz (Section II A) and reviewing the typical formulation of the LM (Section II B). We then show how the cost-scaling for applying the LM to the CJAGP may be reduced (Section II C), how the strong zero variance principle is maintained (Section II D), and how one can avoid constructing the LM matrices when desirable (Section II E). After presenting computational details (Section III A), we then present data on the improved optimization efficiency (Section III B) as well as the accuracy of the method in the triple bond dissociations of N_2 (Section III C) and $[ScO]^+$ (Section III D), before concluding and offering remarks on possible future directions (Section IV).

II. THEORY

A. Basics

In this paper we seek to optimize the CJAGP ansatz,

$$|\Psi\rangle = \exp(\hat{K})|\Phi\rangle, \quad (1)$$

in which the unitary orbital rotation operator $\exp(\hat{K})$ is defined by the anti-Hermitian operator

$$\hat{K} = \sum_{p<q} K_{pq}(a_p^\dagger a_q - a_q^\dagger a_p) \quad (2)$$

and

$$|\Phi\rangle = \exp\left(\sum_{ij} J_{ij}\hat{n}_i\hat{n}_j\right)\left(\sum_{rs} F_{rs}a_r^\dagger a_s^\dagger\right)^{N/2}|0\rangle \quad (3)$$

is the JAGP ansatz with pairing matrix \mathbf{F} and Jastrow factor coefficients \mathbf{J} . In Eq. (3), N is the (even) number of electrons, r and s are restricted to α and β spin-orbitals, respectively, and i and j range over all spin-orbitals. Note that unless otherwise stated, indices in this paper are assumed to range over all spin-orbitals. We will make use of the fermionic creation and destruction operators, a_p^\dagger and a_p , which create or destroy an electron in spin-orbital p and which obey the usual anti-commutation rules. We also employ the number operators $\hat{n}_p = a_p^\dagger a_p$.

The development of improved optimization methods

for the orbital rotation defined by \hat{K} is important because it is this rotation that allows the Jastrow factor to act as a limited CC doubles operator,

$$e^{\hat{K}} e^{\sum_{ij} J_{ij}\hat{n}_i\hat{n}_j} e^{-\hat{K}} = \exp\left(\sum_{ijkl} T_{ij}^{kl} a_k^\dagger a_l^\dagger a_i a_j\right), \quad (4)$$

$$T_{ij}^{kl} = \sum_{pq} U_{ip}^* U_{kp} J_{pq} U_{jq}^* U_{lq}, \quad (5)$$

where \mathbf{U} results from exponentiating the antisymmetrization of the upper-triangular \mathbf{K} [3]. Given the potentially highly multi-reference nature of the geminal power [10], this raises the tantalizing question of whether the CJAGP can act as an effective surrogate for much more complex complete-active-space-based multi-reference CC ansatzes that have outstanding accuracy but steeply scaling computational costs. Although initial investigations into the CJAGP showed promise [3], they were limited by the shortcomings of combining the quasi-Newton L-BFGS method with VMC. We will therefore turn our attention to creating a more effective optimization scheme in order to push CJAGP into larger and more interesting systems.

B. Traditional Linear Method

The LM [6–9] optimization scheme works by solving the Schrödinger equation (SE) in the subspace of Hilbert space spanned by the approximate wave function and its first derivatives with respect to its variables $\boldsymbol{\mu}$, which we write concisely as

$$|\Psi^0\rangle \equiv |\Psi\rangle \quad |\Psi^x\rangle \equiv \frac{\partial|\Psi\rangle}{\partial\mu_x} \quad x \in \{1, 2, \dots, n_v\}. \quad (6)$$

As these functions may not be orthogonal, the SE to be solved is a generalized eigenvalue problem,

$$\mathbf{H}\mathbf{c} = E\mathbf{S}\mathbf{c} \quad (7)$$

$$H_{xy} = \langle\Psi^x|\hat{H}|\Psi^y\rangle \quad \forall \quad x, y \in \{0, 1, 2, \dots, n_v\} \quad (8)$$

$$S_{xy} = \langle\Psi^x|\Psi^y\rangle \quad \forall \quad x, y \in \{0, 1, 2, \dots, n_v\} \quad (9)$$

Assuming the initial wave function is close to the energy minimum, then the ratios c_x/c_0 for $x > 0$ can be expected to be small, as the optimal wave function in the LM subspace should be a small change from $|\Psi\rangle$ (this smallness can be ensured by penalizing the $x > 0$ diagonal elements H_{xx} [8]). Having solved Eq. (7) for \mathbf{c} , we may then update our wave function by a reverse Taylor expansion,

$$|\Psi(\boldsymbol{\mu})\rangle \rightarrow |\Psi(\boldsymbol{\mu} + \mathfrak{O}/c_0)\rangle \approx |\Psi\rangle + \sum_{x=1}^{n_v} \frac{c_x}{c_0} |\Psi^x\rangle, \quad (10)$$

where \mathfrak{c} is the length- n_v vector obtained by removing the first element (c_0) from \mathbf{c} . The key role of Monte Carlo is to evaluate the matrices \mathbf{H} and \mathbf{S} , which is done by a resolution of the identity in terms of occupation number vectors \mathbf{n} (in real space we would instead use an integral over positions) over which a stochastic sample is taken,

$$\begin{aligned} A_{xy} &= \sum_{\mathbf{n}} \langle \Psi^x | \mathbf{n} \rangle \langle \mathbf{n} | \hat{A} | \Psi^y \rangle \\ &= \sum_{\mathbf{n}} |\langle \mathbf{n} | \Psi \rangle|^2 \frac{\langle \Psi^x | \mathbf{n} \rangle}{\langle \Psi | \mathbf{n} \rangle} \frac{\langle \mathbf{n} | \hat{A} | \Psi^y \rangle}{\langle \mathbf{n} | \Psi \rangle} \\ &\approx \sum_{\mathbf{n} \in \xi} \frac{\langle \Psi^x | \mathbf{n} \rangle}{\langle \Psi | \mathbf{n} \rangle} \frac{\langle \mathbf{n} | \hat{A} | \Psi^y \rangle}{\langle \mathbf{n} | \Psi \rangle} \end{aligned} \quad (11)$$

For \mathbf{H} we set $\hat{A} = \hat{H}$ while for \mathbf{S} we set \hat{A} to the identity operator. In this paper the sample of configurations ξ will be drawn from the distribution $|\langle \mathbf{n} | \Psi \rangle|^2$, but any distribution $|Q(\mathbf{n})|^2$ can be used if the right hand side of Eq. (11) is modified to $\sum_{\mathbf{n} \in \xi} \langle \Psi^x | \mathbf{n} \rangle \langle \mathbf{n} | \hat{A} | \Psi^y \rangle / |Q(\mathbf{n})|^2$. Note that the normalization constant for the sampled distribution may be ignored, as it will appear on either side of Eq. (7) and will thus not affect the solution \mathbf{c} . For CJAGP, we will retain the use of Eq. (11) for most but not all elements of \mathbf{H} and \mathbf{S} , as shown in Figure 1.

To see why we do not retain the traditional approach for all matrix elements, consider element H_{xy} in which μ_y is the orbital rotation variable K_{pq} , in which case we must evaluate

$$\langle \mathbf{n} | \hat{H} | \Psi^y \rangle = \frac{\partial \langle \mathbf{n} | \hat{H} | \Psi \rangle}{\partial K_{pq}} = \langle \mathbf{n} | \hat{H} (a_p^\dagger a_q - a_q^\dagger a_p) | \Psi \rangle, \quad (12)$$

in which the two-electron component of \hat{H} combines with the pq -indexed excitations to create triple excitations acting on the configuration \mathbf{n} . While such triple-excitation terms may be evaluated using the same approach as for double excitations (as in the JAGP energy evaluation [5]), the cost scaling for this approach is N^6 , which is much higher than the N^4 scaling that can be achieved [5] when μ_y corresponds to a Jastrow or AGP variable. (Note that to get the LM's overall cost scaling, one must add an additional factor of N if the statistical uncertainty of extensive quantities is to be held constant due to the requisite increase in the sample length.)

C. Lower Scaling Matrix Builds

For a general two-body operator of the form

$$\hat{A} = A_0 + \sum_{pq} A_{pq}^p a_p^\dagger a_q + \sum_{pqrs} A_{rs}^{pq} a_p^\dagger a_q^\dagger a_s a_r \quad (13)$$

and a wave function ansatz consisting of a JAGP augmented by an orbital rotation as in Eq. (1), the per-sample cost scaling to build the matrix \mathbf{A} can be reduced

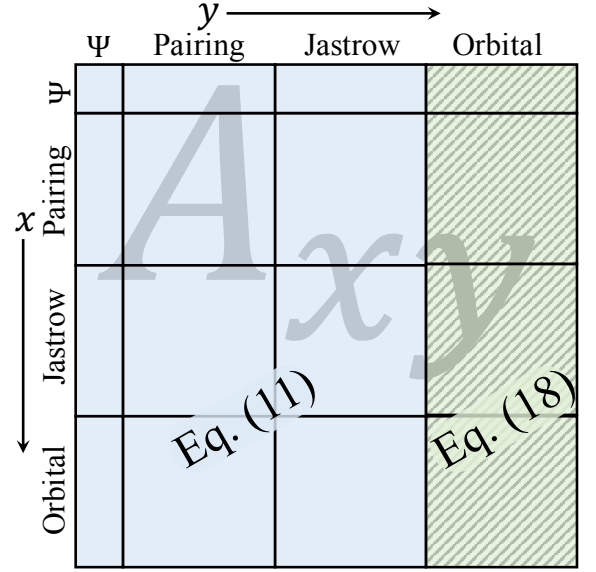


FIG. 1: Equations used for evaluating different subsections of the LM matrices \mathbf{H} and \mathbf{S} .

to N^5 by working in the one-particle basis in which $\hat{\mathcal{K}} = 0$ and by performing the Monte-Carlo-sampled resolution of the identity in a slightly different way. Note that an arbitrary rotation of the one-particle basis (after which \hat{A} will have the same form but different coefficients) can be achieved by converting

$$|\Psi\rangle \rightarrow e^{-\hat{\mathcal{L}}} |\Psi\rangle \quad \hat{A} \rightarrow e^{-\hat{\mathcal{L}}} \hat{A} e^{\hat{\mathcal{L}}} \quad (14)$$

using an anti-Hermitian one-body operator $\hat{\mathcal{L}}$ that defines the rotation. At the end of each LM iteration, at which point $\hat{\mathcal{K}}$ may be nonzero due to the LM update of Eq. (10), we may thus “reset” $\hat{\mathcal{K}}$ to 0 via a basis-rotation with $\hat{\mathcal{L}} = \hat{\mathcal{K}}$. The one- and two-electron coefficients needed to represent \hat{A} in the new basis, i.e. A_p^q and A_{rs}^{pq} in Eq. (13), can be evaluated at an N^5 cost as per a standard atomic-to-molecular-orbital conversion of the one- and two-electron integrals [52]. As the basis rotation is required only once per LM iteration, rather than once per sample, its cost is negligible compared to the sampling effort involved in estimating the matrix \mathbf{A} .

Working in the $\hat{\mathcal{K}} = 0$ one-particle basis, we may express the difficult $\mu_y = K_{pq}$ matrix element as

$$\begin{aligned} A_{xy} &= \langle \Psi^x | \hat{A} | \Psi^y \rangle \\ &= \left[\langle \Psi^x | \frac{\partial}{\partial K_{pq}} (\hat{A} e^{\hat{\mathcal{K}}} | \Phi \rangle) \right]_{\hat{\mathcal{K}}=0} \\ &= \left[\langle \Psi^x | \frac{\partial}{\partial K_{pq}} (e^{\hat{\mathcal{K}}} e^{-\hat{\mathcal{K}}} \hat{A} e^{\hat{\mathcal{K}}} | \Phi \rangle) \right]_{\hat{\mathcal{K}}=0} \\ &= \langle \Psi^x | \hat{\mathcal{C}} | \Phi \rangle + \langle \Psi^x | \hat{D} \hat{A} | \Phi \rangle \\ &= \sum_{\mathbf{n}} \langle \Psi^x | \mathbf{n} \rangle \langle \mathbf{n} | \hat{\mathcal{C}} | \Phi \rangle + \langle \Psi^x | \hat{D} | \mathbf{n} \rangle \langle \mathbf{n} | \hat{A} | \Phi \rangle \end{aligned} \quad (15)$$

where we have defined

$$\hat{C} \equiv \left[\frac{\partial(e^{-\hat{K}} \hat{A} e^{\hat{K}})}{\partial K_{pq}} \right]_{\hat{K}=0} = [\hat{A}, a_p^+ a_q - a_q^+ a_p] \quad (16)$$

$$\hat{D} \equiv \left[\frac{\partial e^{\hat{K}}}{\partial K_{pq}} \right]_{\hat{K}=0} = a_p^+ a_q - a_q^+ a_p \quad (17)$$

The rationale for these placements of the identity resolutions is that they isolate the difficult operators \hat{A} and \hat{C} such that no term involves more than a double excitation on $|\mathbf{n}\rangle$ (the uncontracted triple excitations in the commutator of \hat{C} cancel each other as usual), thus avoiding the triple excitation in Eq. (12) that led to N^6 scaling. Having placed our identity resolutions, we may now evaluate them stochastically on a sample ξ drawn from $|\langle \Phi | \mathbf{n} \rangle|^2$ in order to produce our Monte Carlo estimate of the matrix element:

$$A_{xy} \approx \sum_{\mathbf{n} \in \xi} \frac{\langle \Psi^x | \mathbf{n} \rangle}{\langle \Phi | \mathbf{n} \rangle} \frac{\langle \mathbf{n} | \hat{C} | \Phi \rangle}{\langle \mathbf{n} | \Phi \rangle} + \frac{\langle \Psi^x | \hat{D} | \mathbf{n} \rangle}{\langle \Phi | \mathbf{n} \rangle} \frac{\langle \mathbf{n} | \hat{A} | \Phi \rangle}{\langle \mathbf{n} | \Phi \rangle} \quad (18)$$

It now remains to evaluate these matrix element estimates for the identity and Hamiltonian operators involved in the LM.

For the overlap matrix \mathbf{S} , for which \hat{A} is the identity, \hat{C} vanishes and Eq. (18) simplifies to

$$S_{xy} \approx \sum_{\mathbf{n} \in \xi} \frac{\langle \Psi^x | (a_p^+ a_q - a_q^+ a_p) | \mathbf{n} \rangle}{\langle \Phi | \mathbf{n} \rangle}. \quad (19)$$

As shown in Appendix A, the per-sample cost to evaluate these $\mu_y = K_{pq}$ matrix blocks (i.e. the Eq. (18) blocks for \mathbf{S} in Figure 1) grows as only N^4 .

For the Hamiltonian matrix \mathbf{H} , for which $\hat{A} = \hat{H}$, things are not so simple, although it is possible to avoid an N^6 per-sample cost scaling. To begin, we may recognize that the right hand part Eq. (18) becomes a simple modification of Eq. (19) in which each term is scaled by the JAGP local energy $\langle \mathbf{n} | \hat{H} | \Phi \rangle / \langle \mathbf{n} | \Phi \rangle$ (which can be evaluated at an N^4 per-sample cost [5]), and so its contribution to \mathbf{H} can be evaluated at an N^4 per-sample cost by a direct analogue of the approach for \mathbf{S} given in Appendix A. In the left-hand part of Eq. (18), consider first the derivative ratios

$$\mathcal{D}_{\mathbf{n}}(\mu_x) \equiv \frac{\langle \Psi^x | \mathbf{n} \rangle}{\langle \Phi | \mathbf{n} \rangle}. \quad (20)$$

For μ_x either a pairing matrix element or a Jastrow coefficient, these ratios have been evaluated previously for the JAGP [5]. When μ_x is an orbital rotation variable K_{pq} , the ratios are

$$\mathcal{D}_{\mathbf{n}}(K_{pq}) = \frac{\langle \Phi | (a_p^+ a_q - a_q^+ a_p) | \mathbf{n} \rangle}{\langle \Phi | \mathbf{n} \rangle}, \quad (21)$$

which can be evaluated efficiently as shown in Appendix A, specifically in Eq. (A3).

The final term needed to construct \mathbf{H} , and the one responsible for the overall N^5 per-sample cost scaling of the construction, is the term in Eq. (18) containing \hat{C} . We will worry only about the two-electron component of \hat{H} (the reader may convince herself that the one-electron component is less expensive), for which we must evaluate

$$\frac{1}{\langle \mathbf{n} | \Phi \rangle} \langle \mathbf{n} | \left[\sum_{ijkl} g_{kl}^{ij} a_i^+ a_j^+ a_l a_k, a_p^+ a_q - a_q^+ a_p \right] | \Phi \rangle \quad (22)$$

where g_{kl}^{ij} are the usual two-electron integrals [52]. Defining the double excitation ratios

$$Q_{kl}^{ij} \equiv \frac{\langle \mathbf{n} | a_i^+ a_j^+ a_l a_k | \Phi \rangle}{\langle \mathbf{n} | \Phi \rangle}, \quad (23)$$

which are derivatives of the JAGP local energy with respect to g_{kl}^{ij} (see Eq. (34) of Ref. [5]) and can thus all be evaluated for the same N^4 cost-per-sample scaling as the local energy itself, one may expand Eq. (22) as

$$\sum_{ijk} \left(g_{pk}^{ij} Q_{qk}^{ij} + g_{kp}^{ij} Q_{kq}^{ij} - g_{jk}^{iq} Q_{jk}^{ip} - g_{jk}^{qi} Q_{jk}^{pi} + g_{pk}^{ij} Q_{ij}^{qk} + g_{kp}^{ij} Q_{ij}^{kq} - g_{jk}^{iq} Q_{ip}^{jk} - g_{jk}^{qi} Q_{pi}^{jk} \right). \quad (24)$$

Each of these terms can clearly be evaluated for a per-sample cost scaling as N^5 , giving the explicit construction of the CJAGP \mathbf{H} matrix according to the scheme in Figure 1 an overall per-sample cost that scales as N^5 . This is better than the N^6 per-sample cost resulting from a naive application of the traditional LM matrix build, but nonetheless a higher scaling than for JAGP.

D. Strong Zero Variance

In the traditional LM, the stochastic approximation to the generalized eigenvalue problem in Eq. (7) has the important property of satisfying what is known as the strong zero variance principle (SZVP), which says that the solution of the eigenproblem will have no statistical uncertainty if the exact wave function exists within the span of the current wave function and its first derivatives. In practice this means that as an accurate wave function is approached, statistical uncertainty in the LM is greatly reduced. This is a generalization of the standard VMC zero variance principle, in which the energy has no uncertainty if the wave function ansatz itself is exact. To see where the SZVP comes from, consider the following rearrangement of Eq. (7) in which the matrices have been approximated stochastically as in the traditional LM (i.e.

via Eq. (11))

$$0 = \sum_{y=0}^{n_v} (H_{xy} - E S_{xy}) c_y \quad (25)$$

$$\approx \sum_{y=0}^{n_v} \sum_{\mathbf{n} \in \xi} \frac{\langle \Psi^x | \mathbf{n} \rangle \langle \mathbf{n} | (\hat{H} - E) | \Psi^y \rangle}{\langle \Phi | \mathbf{n} \rangle \langle \mathbf{n} | \Phi \rangle} c_y \quad (26)$$

$$= \sum_{\mathbf{n} \in \xi} \frac{\langle \Psi^x | \mathbf{n} \rangle \langle \mathbf{n} | (\hat{H} - E) \sum_{y=0}^{n_v} | \Psi^y \rangle c_y}{\langle \Phi | \mathbf{n} \rangle \langle \mathbf{n} | \Phi \rangle} \quad (27)$$

If the exact wave function exists within the LM subspace, which is to say there is a vector \mathbf{c} such that

$$(\hat{H} - E) \sum_{y=0}^{n_v} | \Psi^y \rangle c_y = 0, \quad (28)$$

then the terms in Eq. (27) vanish independently for every \mathbf{n} , and so the exact energy E and the vector \mathbf{c} giving the exact wave function will be found during the diagonalization of Eq. (7) regardless of which random sample ξ is taken. In other words, they will be found with zero variance.

Although the present approach does not satisfy the SZVP exactly, its deviation from the SZVP vanishes quadratically as the exact wave function is approached. To see why, replace \mathbf{H} and \mathbf{S} with Figure 1's stochastic approximations and (without loss of generality) set $c_0 = 1$, at which point the deviation of Eq. (25) from zero (i.e. the deviation from the SZVP) becomes

$$\begin{aligned} \eta_x &= \sum_{\mathbf{n} \in \xi} \frac{1}{|\langle \Phi | \mathbf{n} \rangle|^2} \left[\langle \Psi^x | \mathbf{n} \rangle \langle \mathbf{n} | (\hat{H} - E) | \Psi \rangle + \right. \\ &\quad \left. \sum_{y=1}^{n_v} \langle \Psi^x | \frac{\partial}{\partial \mu_y} \left(e^{\hat{\mathcal{K}}} | \mathbf{n} \rangle \langle \mathbf{n} | e^{-\hat{\mathcal{K}}} (\hat{H} - E) | \Psi \rangle \right) c_y \right]_{\hat{\mathcal{K}}=0} \\ &= \sum_{\mathbf{n} \in \xi} \frac{1}{|\langle \Phi | \mathbf{n} \rangle|^2} \left[\langle \Psi^x | \mathbf{n} \rangle \langle \mathbf{n} | (\hat{H} - E) \sum_{y=0}^{n_v} | \Psi^y \rangle c_y + \right. \\ &\quad \left. \sum_{y=1}^{n_v} \langle \Psi^x | \frac{\partial}{\partial \mu_y} \left(e^{\hat{\mathcal{K}}} | \mathbf{n} \rangle \langle \mathbf{n} | e^{-\hat{\mathcal{K}}} \right) (\hat{H} - E) | \Psi \rangle c_y \right]_{\hat{\mathcal{K}}=0} \end{aligned}$$

If we again assume that the (un-normalized) exact wave function $|\Psi_0\rangle = |\Psi\rangle + \sum_{z=1}^{n_v} |\Psi^z\rangle c_z$ exists in the LM subspace, which implies that

$$(\hat{H} - E) | \Psi \rangle = - \sum_{z=1}^{n_v} (\hat{H} - E) | \Psi^z \rangle c_z, \quad (29)$$

then the deviation from the SZVP simplifies to

$$\eta_x = - \sum_{y=1}^{n_v} \sum_{z=1}^{n_v} c_y c_z \sum_{\mathbf{n} \in \xi} \frac{Q_{xyz}^{(\mathbf{n})}}{|\langle \Phi | \mathbf{n} \rangle|^2} \quad (30)$$

$$Q_{xyz}^{(\mathbf{n})} \equiv \left[\langle \Psi^x | \frac{\partial}{\partial \mu_y} \left(e^{\hat{\mathcal{K}}} | \mathbf{n} \rangle \langle \mathbf{n} | e^{-\hat{\mathcal{K}}} \right) (\hat{H} - E) | \Psi^z \rangle \right]_{\hat{\mathcal{K}}=0}$$

Thus the deviation from the SZVP vanishes quadratically as $|\mathfrak{D}|^2$ with the difference \mathfrak{D} between the current and exact wave functions. This is in stark contrast to the previous quasi-Newton optimization strategy [3] which lacked any kind of zero variance principle for the optimization updates, a fact that likely explains our previous observation that greatly increased sample lengths compared to the traditional LM were needed to stabilize the quasi-Newton approach.

Note that while it is possible to approximate CJAGP's \mathbf{S} matrix at an N^4 per-sample cost using the traditional LM's stochastic approach of Eq. (11), doing so would violate even this quadratic approach to the SZVP when used together with Figure 1's approximation for \mathbf{H} . Indeed, we have observed that drastically larger sample sizes are required when one mixes the traditional method for approximating \mathbf{S} with our new method for approximating \mathbf{H} , and so we also approximate \mathbf{S} via Figure 1 for the sake of reducing statistical uncertainty, even though this approximation is more complicated.

E. Avoiding Matrix Builds

Although the LM typically works by first building the matrices \mathbf{H} and \mathbf{S} and then solving the generalized eigenvalue problem of Eq. (7), Krylov subspace (KS) methods [53] such as the Davidson [54] or Arnoldi [55] methods can be employed to eschew the matrix builds altogether. Such a KS approach has been used previously [56] in the context of stochastic reconfiguration [35, 57], and here we give some details for how such approaches can be generalized to both the traditional LM and the newly proposed variant for CJAGP. Instead of requiring the matrices to be built, KS methods typically only require the ability to operate the matrix on an arbitrary vector, which in the context of either the LM or stochastic reconfiguration can be advantageous when the number of wave function variables n_v becomes large.

In the traditional LM, a KS method will require evaluation of matrix-vector products $\mathbf{A}\mathbf{c}$ with the stochastic matrix approximation given in Eq. (11):

$$\sum_y A_{xy} c_y \approx \sum_{\mathbf{n} \in \xi} \frac{\langle \Psi^x | \mathbf{n} \rangle}{\langle \Psi | \mathbf{n} \rangle} \sum_y \frac{\langle \mathbf{n} | \hat{A} | \Psi^y \rangle}{\langle \mathbf{n} | \Psi \rangle} c_y \quad (31)$$

For wave functions like the JAGP [5] for which the derivative vectors $\langle \Psi^x | \mathbf{n} \rangle / \langle \Psi | \mathbf{n} \rangle$ and $\langle \mathbf{n} | \hat{A} | \Psi^y \rangle / \langle \mathbf{n} | \Psi \rangle$ can be evaluated efficiently, each sample's contribution to the overall matrix vector product can be computed via a simple dot product. If, for example, storing or communicating the matrix would be prohibitive, this approach offers a lower-memory, lower-communication alternative.

For the approach proposed here for the CJAGP ansatz,

the matrix vector product takes on two parts. For the portion of the sum over y covering the non-differentiated term, the pairing matrix derivatives, and the Jastrow derivatives, the evaluation is the same as in Eq. (31). For the portion of the sum in which y runs over orbital rotation variables K_{pq} , we use Eq. (18) to write

$$\sum_{y \in \text{orb. rot.}} A_{xy} c_y \approx \sum_{\mathbf{n} \in \xi} \frac{\langle \Psi^x | \mathbf{n} \rangle \langle \mathbf{n} | \check{A} | \Phi \rangle}{\langle \Phi | \mathbf{n} \rangle \langle \mathbf{n} | \Phi \rangle} + \frac{\langle \Psi^x | \check{B} | \mathbf{n} \rangle \langle \mathbf{n} | \hat{A} | \Phi \rangle}{\langle \Phi | \mathbf{n} \rangle \langle \mathbf{n} | \Phi \rangle} \quad (32)$$

$$\check{B} \equiv \sum_{p < q} c_{pq} (a_p^+ a_q - a_q^+ a_p) \quad (33)$$

$$\check{A} \equiv [\hat{A}, \check{B}] \quad (34)$$

In the definition of \check{B} we have relabeled the sum on y over orbital rotations by the orbital indices p and q that label the individual orbital rotation variables. Crucially, because \check{B} is a one-electron operator, \check{A} is a two-electron operator with exactly the same form as \hat{A} . Moreover, the coefficients for \check{A} are independent of \mathbf{n} and can thus be precomputed at an N^5 cost *before* the sample is taken, so that the actual per-sample cost of evaluating the first term in Eq. (32) scales as only N^4 . The second term in Eq. (32) also has a per-sample cost scaling as N^4 , as it amounts to a weighted sum over the matrix elements of Eq. (19) scaled either by one (if $\mathbf{A} = \mathbf{S}$) or by the JAGP local energy (if $\mathbf{A} = \mathbf{H}$). Thus we see that in contrast to building the CJAGP LM matrices, which due to Eq. (24) has a per-sample cost scaling of N^5 , operating these matrices on an arbitrary vector without building them has a per-sample cost scaling of only N^4 . In large systems this reduced scaling could be an advantage, depending on how many matrix vector products are required for the chosen Krylov subspace method. Systems studied in this work are too small for this reduced scaling to be beneficial, but we present the option of avoiding matrix builds anyways as it should be useful in future work.

III. RESULTS

A. Computational Details

CJAGP results were obtained using our own software for VMC in Hilbert space, with one- and two-electron integrals for the Hamiltonian taken from Psi3 [58]. Complete-active space self-consistent field (CASSCF) [59, 60], full configuration interaction (FCI) [61, 62], complete-active space second order perturbation theory (CASPT2) [63], and size-consistency-corrected multi-reference configuration interaction (MRCI+Q) [64, 65] results were obtained with Molpro [66]. Except for the (6e,12o) CASSCF result displayed in Figure 5, all other cases of CASSCF, CASPT2, and MRCI+Q employed a

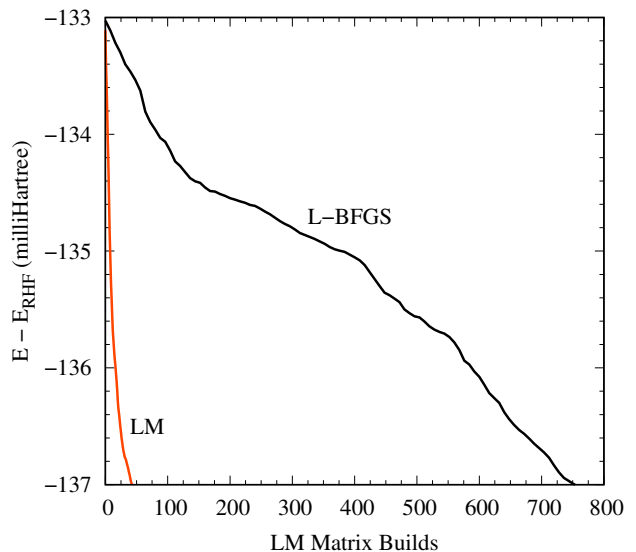


FIG. 2: Convergence of the last few mE_h of correlation energy for the LM and L-BFGS optimization approaches for H_2O in a 6-31G basis with $r_{OH} = 1.0 \text{ \AA}$ and $\angle HOH = 109.57^\circ$, plotted against the number of LM matrix builds completed after passing a correlation energy of $-133 mE_h$. The converged CJAGP correlation energy is $-137.3 mE_h$. See Section III B for further details.

minimal (6e,6o) active space containing the three pairs of bonding/antibonding orbitals for the triple bonds of N_2 and $[ScO]^+$. Results for restricted and unrestricted Hartree Fock (RHF and UHF) [67] and coupled cluster with singles, doubles, and perturbative triples (CCSD(T) and UCCSD(T)) [4] were obtained with QChem [68, 69]. A 6-31G [70] basis was used in all cases, and post-CASSCF methods (as well as CJAGP) froze the N 1s, O 1s, and Sc 1s, 2s, and 2p orbitals.

In the optimization of our CJAGP wave function, some constraints were placed on the wave function to improve the ease of convergence. For the Jastrow factor, the coefficients were constrained to be symmetric between α and β electrons, so $J_{i\alpha j\alpha} = J_{i\beta j\beta}$ and $J_{i\alpha j\beta} = J_{i\beta j\alpha}$. For the pairing matrix, we constrained \mathbf{F} to be symmetric, resulting in an AGP reference with singlet spin. Finally, we added further constraints to ensure Jastrow coefficients and pairing matrix elements that should be equal by molecular symmetry were indeed equal. For example, in N_2 and $[ScO]^+$ the Jastrow coefficients for equivalent s- p_x and s- p_y couplings were constrained to be equal.

Note that the optimized CJAGP energy has two possible sources of statistical uncertainty. First, there is the usual uncertainty when estimating the final wave function's energy using VMC. Second, statistical uncertainty in the LM update direction \mathfrak{s} prevents the optimal variable values from being found precisely. In practice we observe the latter effect to be dominant, making the estimation of the overall method's statistical uncertainty somewhat difficult, as we do not wish to run a large number of separate optimizations at each molecular geometry

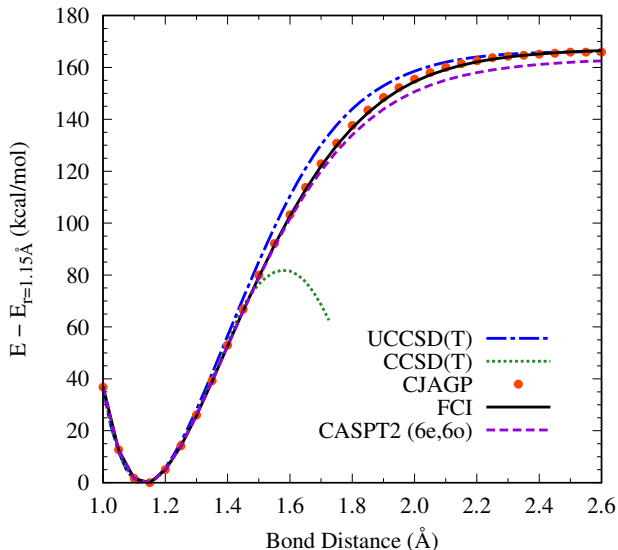


FIG. 3: Potential energy curves for N_2 dissociation in a 6-31G basis, with each curve shifted so that the zero of energy occurs at 1.15 \AA . See Section III C for further details.

to collect statistics. Instead, we have fit CJAGP’s energy error over the dissociation curves to a smooth third order polynomial (e.g. see Figure 4) and then estimated the statistical uncertainty of the energies based on the deviations of the actual points from this smooth curve. Assuming these deviations are normally distributed, we find 95% confidence intervals of ± 0.13 kcal/mol in N_2 and ± 0.3 kcal/mol in $[\text{ScO}]^+$.

B. Convergence

Figure 2 shows, in H_2O near equilibrium, the convergence of the present LM approach compared to the previous [3] quasi-Newton L-BFGS optimization scheme. The L-BFGS approach’s idea was to optimize the orbital rotation \hat{K} on a surface $E(\mathbf{K})$ on which the other variables took on their optimal values (i.e. the Jastrow and pairing variables for a given \mathbf{K} were taken as those that minimized the energy for that \mathbf{K}). In practice this surface was achieved by using the LM to reoptimize the Jastrow and pairing variables at each L-BFGS step, and so we are able to compare the number of LM matrix builds required in that scheme to the number required by the present full LM approach. While this comparison is somewhat imperfect as the previous LM matrix builds were less expensive than the present ones that also include the orbital rotation variables, Figure 2 nonetheless displays the stark contrast in optimization efficiency between the two approaches.

Note that this example was carried out under what might be called “exact sampling” (meaning that each configuration \mathbf{n} was visited exactly once and its contribution to averages scaled by the wave function weight $|\langle \mathbf{n} | \Psi \rangle|^2$)

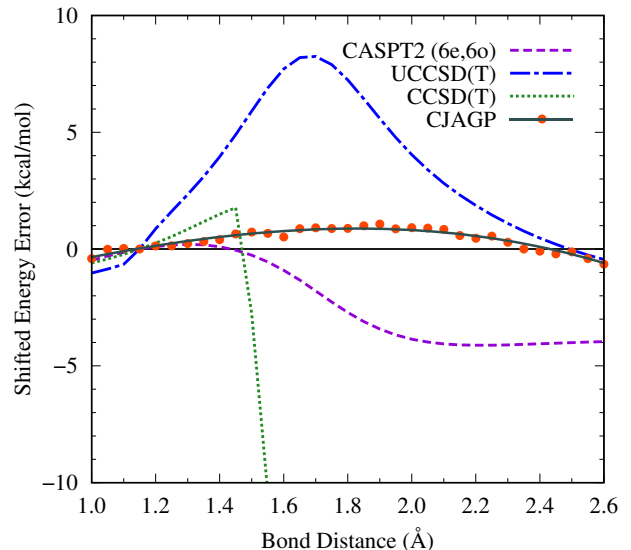


FIG. 4: Energy deviations from FCI during N_2 dissociation in a 6-31G basis, with each curve shifted by a constant so that it crosses zero at a bond distance of 1.15 \AA . For CJAGP, the line is a cubic polynomial fit to the points to give a sense of statistical uncertainty. See Section III C for further details.

so that statistical uncertainty was not present. In addition to being useful for debugging, such sampling allows us to test whether, independent of stochastic issues, the present LM outperforms L-BFGS, and indeed it clearly does. Note that the comparison becomes even more favorable for the present LM approach when a stochastic Markov-chain-based sample is used, as the LM obeys a zero variance principle while the L-BFGS approach does not. In practice, we therefore see that not only is the LM a superior optimization method, but that its inherently lower statistical uncertainty allows it to operate effectively with much smaller sample sizes than are required to stabilize the previous L-BFGS approach. One way to emphasize this advantage is to point out that in our previous study [3], the stochastic sample sizes needed to stabilize the L-BFGS method were in all cases larger than the Hilbert spaces themselves (note that this is not uncommon for stochastic approaches in small systems), whereas the present study’s sample lengths of 1.6×10^7 for N_2 and 2.56×10^7 for $[\text{ScO}]^+$ were in both cases smaller than the Hilbert spaces in question.

C. N_2

The dissociation of the nitrogen dimer’s triple bond has long been used as a benchmark for multi-reference methods in quantum chemistry. As was seen previously [3] in H_2O and HF, the limited CC-like nature of the CJAGP’s orbital-rotated Jastrow factor appears to capture a large fraction of the dynamic correlation energy while maintaining the ability to help capture static correlation in

TABLE I: Energies for the N_2 stretch in a 6-31G basis. FCI is reported in E_h , with other methods reported as the difference from FCI in mE_h . The last row gives the non-parallelity errors in mE_h . See Section III C for further details.

R (Å)	FCI	RHF	UHF	CCSD(T)	UCCSD(T)	CASSCF	CASPT2	CJAGP
1.00	-109.0467	211.4	211.4	-0.8	-0.8	85.5	13.7	6.8
1.05	-109.0857	223.3	223.3	-0.5	-0.5	86.5	14.0	7.4
1.10	-109.1034	235.7	235.7	-0.2	-0.2	87.4	14.3	7.5
1.15	-109.1059	248.8	247.5	0.2	0.9	88.3	14.5	7.4
1.20	-109.0981	262.4	252.9	0.6	2.3	89.2	14.7	7.6
1.25	-109.0835	276.6	252.9	1.1	3.5	90.0	14.8	7.6
1.30	-109.0648	291.3	249.1	1.5	4.6	90.8	14.8	7.8
1.35	-109.0438	306.6	242.7	2.1	5.8	91.6	14.8	7.9
1.40	-109.0221	322.4	234.6	2.6	7.2	92.3	14.6	8.1
1.45	-109.0005	338.8	225.3	3.0	8.7	93.0	14.4	8.5
1.50	-108.9797	352.9	215.1	-4.2	10.3	93.5	14.1	8.6
1.55	-108.9602	363.1	203.9	-16.3	11.9	93.9	13.6	8.5
1.60	-108.9423	370.5	191.9	-33.6	13.2	94.0	13.0	8.3
1.65	-108.9260	376.2	179.1	-56.4	13.9	94.0	12.4	8.8
1.70	-108.9116	380.9	166.2	-84.9	14.0	93.6	11.6	8.9
1.75	-108.8989	385.2	153.5		13.5	92.9	10.9	8.8
1.80	-108.8880	389.6	141.5		12.4	91.9	10.2	8.8
1.85	-108.8788	394.3	130.6		11.1	90.7	9.5	9.0
1.90	-108.8711	399.5	120.9		9.8	89.3	9.0	9.1
1.95	-108.8648	405.2	112.5		8.5	87.7	8.6	8.8
2.00	-108.8597	411.4	105.2		7.3	86.2	8.3	8.9
2.05	-108.8556	418.0	99.1		6.3	84.8	8.1	8.9
2.10	-108.8523	424.9	93.9		5.4	83.4	8.0	8.8
2.15	-108.8496	432.1	89.6		4.6	82.1	8.0	8.3
2.20	-108.8476	439.4	86.1		3.9	81.0	7.9	8.2
2.25	-108.8459	446.7	83.1		3.2	80.0	7.9	8.3
2.30	-108.8446	454.1	80.6		2.6	79.2	8.0	7.9
2.35	-108.8435	461.5	78.6		2.1	78.5	8.0	7.4
2.40	-108.8427	468.7	76.9		1.6	77.8	8.0	7.3
2.45	-108.8420	475.8	75.5		1.2	77.3	8.1	7.1
2.50	-108.8414	482.7	74.3		0.8	76.9	8.1	7.2
2.55	-108.8410	489.5	73.4		0.5	76.5	8.1	6.8
2.60	-108.8406	496.1	72.5		0.2	76.1	8.2	6.4
NPE	N/A	284.6	180.3	87.9	14.8	17.9	6.9	2.7

conjunction with the geminal power [10]. As we see in the N_2 results (Figures 3 and 4 and Table I) these features allow CJAGP to vastly outperform single-reference methods like CCSD(T) and UCCSD(T). Here the catastrophic failure of CCSD(T) may be attributed both to the poor quality of its RHF reference (whose instabilities towards spatial symmetry breaking are responsible for the kink in its potential curve) and to the tendency of spurious interactions between its singlet and triplet amplitude channels to overcorrelate in the strongly correlated regime [71]. Note that the issue of spatial symmetry breaking in the RHF might be avoided by enforcing spa-

tial symmetry throughout the dissociation, but for N_2 we have chosen to present the CCSD(T) results for the minimum energy RHF reference as found via stability analyses [72]. In contrast, CJAGP avoids these issues thanks to its more flexible reference function and the variational nature of its evaluation, which guarantees that spurious couplings between its cluster amplitudes cannot lead to an overcorrelation catastrophe.

More significantly, CJAGP outperforms CASPT2, one of the most affordable and most commonly used multi-reference methods in quantum chemistry. Both its absolute and relative energies show improvements compared

TABLE II: Energies for the $[\text{ScO}]^+$ stretch in a 6-31G basis. MRCI+Q is reported in E_h , with other methods reported as the difference from MRCI+Q in mE_h . The last row gives the non-parallelity errors in mE_h . See Section IIID for further details.

R (Å)	MRCI+Q	RHF	UHF	CCSD(T)	UCCSD(T)	CASSCF (6e,6o)	CASPT2	CJAGP
1.5	-834.6354	345.5	345.5	0.8	0.8	202.9	25.5	69.4
1.6	-834.6631	356.3	356.3	0.6	0.6	206.0	26.0	69.6
1.7	-834.6688	367.5	367.5	0.4	0.4	209.3	26.6	69.0
1.8	-834.6607	379.6	357.9	0.1	8.9	212.8	27.1	68.3
1.9	-834.6445	392.9	340.5	-0.4	7.6	216.4	27.4	68.7
2.0	-834.6243	407.0	324.2	-0.9	5.3	219.9	27.5	69.5
2.1	-834.6024	420.6	310.1	4.7	5.3	223.1	27.1	68.5
2.2	-834.5806	427.4	298.7	14.4	8.7	225.7	26.1	69.3
2.3	-834.5600	429.1	289.8	14.2	12.5	227.3	24.2	70.5
2.4	-834.5413	427.9	281.8	7.6	15.2	228.3	21.6	69.7
2.5	-834.5248	425.2	266.5	-6.3	14.2	228.7	17.0	70.4
2.6	-834.5106	421.6	252.7	-34.4	12.0	228.2		71.6
NPE	N/A	83.5	114.7	48.8	14.9	25.7	10.5	3.3

to those of CASPT2, with the relative energies being particularly accurate: the non-parallelity error (NPE, the difference between the highest and lowest deviations) relative to FCI is less than 2 kcal/mol and less than half that of CASPT2. These improvements are especially significant when one considers that CASPT2's cost scales exponentially due to its complete active space reference, while CJAGP's cost scales only polynomially.

In light of the Jastrow factor's CC-like form and the geminal power's multi-reference nature, it is interesting to compare CJAGP to the performance one might expect from the ideal of a variational singles-and-doubles CC method based on a complete active space self consistent field (CASSCF) wave function reference. As such a theory should outperform even MRCI+Q, one would expect absolute accuracies to be within 1 or 2 mE_h of FCI (see e.g. [73]). Unsurprisingly, given that both its cluster operator and its AGP reference function are more constrained than this ideal, CJAGP does not achieve such accuracies in the absolute energy. Its relative energies are nonetheless quite accurate, suggesting that the missing details that would account for the last few percent of the correlation energy are being left out consistently at all geometries. If supplied with a trial function as accurate as CJAGP, diffusion Monte Carlo [74] would be well placed to capture these final details. One very interesting question going forward is thus whether a real-space Jastrow factor can be devised to replicate the CC qualities of the orbital-rotated Hilbert-space Jastrow.

D. ScO Cation

Due to the importance of transition metals in catalysis and materials science, and the tendency of metal-oxygen bonds to exhibit strong electron correlations, the-

oretical approaches that can deal successfully with such correlations are a high priority. As an initial foray into this regime, we have tested CJAGP on the triple-bond dissociation of $[\text{ScO}]^+$. At first glance, this cation appears quite similar to N_2 in that it also contains one σ and two π bonds. In practice, however, its dissociation is even more fraught, with UCCSD(T) becoming qualitatively unreliable and minimal-active-space CASPT2 exhibiting intruder state problems. As seen in Figure 5, CCSD(T) exhibits its typical failure during multiple bond stretching. UCCSD(T) fares little better, being beset by a Coulson-Fischer point cusp near equilibrium where RHF and UHF separate as well as multiple low-

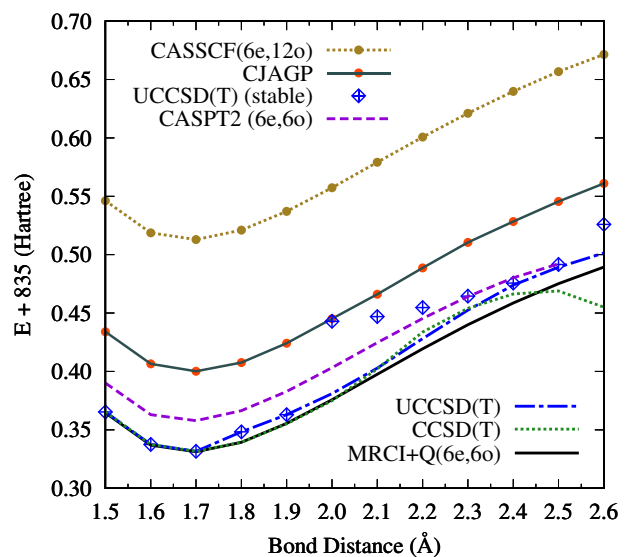


FIG. 5: Total energies during the dissociation of $[\text{ScO}]^+$ in a 6-31G basis. See Section IIID for further details.

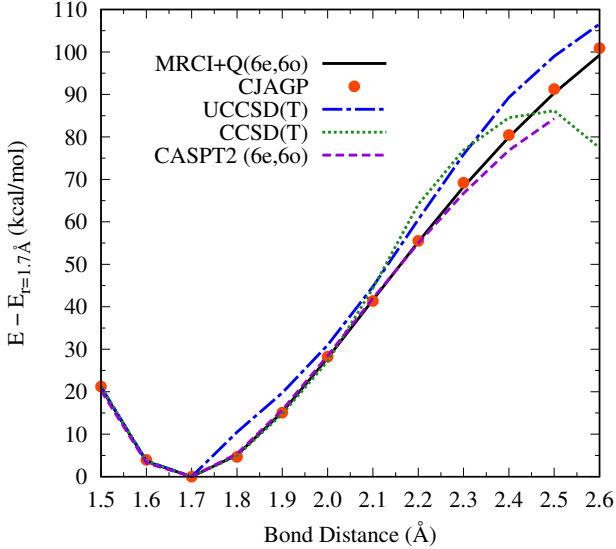


FIG. 6: Energy curves during the dissociation of $[\text{ScO}]^+$ in a 6-31G basis, with each curve shifted so that the zero of energy occurs at 1.7 Å. See Section IIID for further details.

lying UHF determinants as the bond is stretched. If one uses stability analyses to ensure that UCCSD(T) is always based on the lowest energy UHF solution, the result is a UCCSD(T) curve (UCCSD(T) (stable)) with multiple discontinuities. These discontinuities can be avoided by always using the UHF solution with character most similar to the $R = 1.9$ Å UHF state, as we have done for the data labeled UCCSD(T) in Figures 5-7 and Table II, but even in this case UCCSD(T) displays a NPE of 9.3 kcal/mol. One should bear in mind that without benchmark results it would be difficult to know whether this UHF determinant or the lower energy determinants found through stability analyses were the more reasonable starting points, and so it is hard to recommend the use of UCCSD(T) for predicting energy profiles when stretching transition-metal-oxide bonds.

When based on the triple-bond's minimal (6e,6o) active space, CASPT2 proves more reliable than CC and achieves a smaller 6.6 kcal/mol NPE. However, this CASPT2 approach failed to converge at $R = 2.6$ Å due to the presence of an intruder state. One could overcome this problem with either a larger-than-minimal active space or through the use of level shifts [75], but the former may become untenable in larger transition metal systems while the latter introduces an uncontrolled free parameter.

As in N_2 , the active-space-free CJAGP improves on the relative energy of CASPT2 with a NPE of just 2.1 kcal/mol, as seen in Figures 6 and 7. However, as seen in Figure 5 and Table II, the absolute energy errors for CJAGP are now much larger than they were in N_2 . While Figure 5 reveals that CJAGP recovers significantly more correlation energy than even a full valence (6e,12o) CASSCF approach, it is still missing roughly 70 mE_h rel-

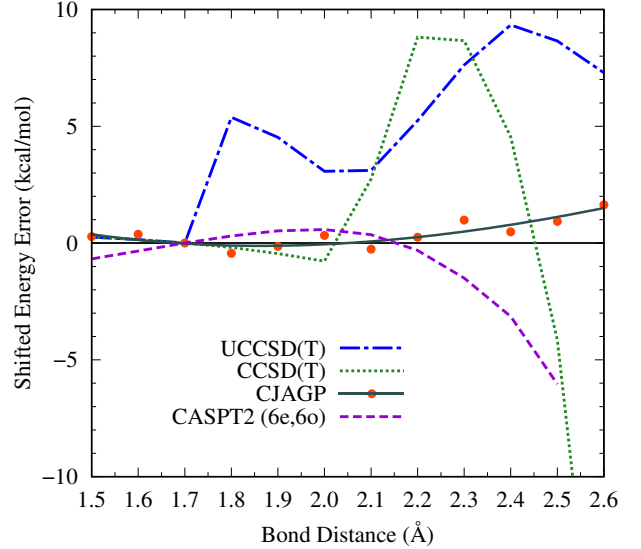


FIG. 7: Energy deviations from MRCI+Q (6e,6o) during $[\text{ScO}]^+$ dissociation in a 6-31G basis, with each curve shifted by a constant so that it crosses zero at a bond distance of 1.7 Å. For CJAGP, the line is a cubic polynomial fit to the points to give a sense of statistical uncertainty. See Section IIID for further details.

ative to the benchmark MRCI+Q. As discussed in Section IIIC, this performance is inferior to what one would expect from a (currently non-existent) CASSCF-based variational CC. Given the excellent shape of CJAGP's potential energy curve (again, NPE is only 2.1 kcal/mol), we do not think the issue lies with the multi-reference Jastrow-AGP combination but instead suspect the missing correlation energy is due to the limited flexibility of the Jastrow operator's CC form (Eq. (5)) when compared to a full CC doubles operator. In other words, we suspect that the limited CC flexibility leads to a limited dynamic correlation recovery, although one that is surprising well balanced across different geometries. As for N_2 , these results strongly suggest that excellent accuracies could be achieved if DMC could use a trial function of CJAGP quality, as DMC is excellent at recovering dynamic correlation details when supplied with a qualitatively correct trial function [9]. Indeed, based on its energy results, CJAGP should be an even better DMC trial function than full-valence CASSCF, and so we feel further motivated to investigate this exciting possibility.

As a final note, we would like to point out that beyond 2.6 Å, the CJAGP optimization failed to converge to a good singlet, likely because at around this geometry a singlet-triplet crossing occurs and the singlet is no longer the ground state [76]. While we hope to investigate CJAGP's prospects for the direct, variational targeting of excited states [77] in the future, we have limited ourselves here to bond distances below 2.6 Å for which the singlet is the ground state.

IV. CONCLUSIONS

We have presented an improved LM optimization scheme for the CJAGP ansatz that achieves an N^5 per-sample cost scaling that drops to N^4 if Krylov subspace methods are employed. This LM optimization obeys the strong zero variance principle in a quadratic sense, and is thus vastly more statistically efficient than the previously employed quasi-Newton approach. In practice this improved optimization scheme has led to drastic reductions in the sample sizes and optimization steps required for variational energy minimization. The key theoretical development facilitating these improvements was the use of an alternative stochastic resolution of the identity in the estimation of the LM matrices or matrix vector products.

With this improved optimization scheme, we showed that CJAGP is vastly more reliable than traditional single-reference CC in two challenging triple-bond dissociations, one involving a transition metal. Further, we showed that for relative energies, the polynomial-cost, active-space-free CJAGP also outperformed the exponentially scaling, active-space-based CASPT2 method. In both examples, the CJAGP relative energies were substantially more accurate than its absolute energy, suggesting to us that the limited flexibility of its cluster operator ($O(N^2)$ variables vs the traditional $O(N^4)$) prevented the capture of the finer details of dynamic correlation in a way that was well balanced across different geometries.

Our findings in this study suggest two important avenues for future investigation. First, given that CJAGP appears to be a better trial function starting point than even a full-valence CASSCF reference, it would be highly desirable to combine it with diffusion Monte Carlo. This is not entirely trivial given that currently the CJ operator exists only in Hilbert (rather than real) space, but we look forward to investigating how its success may inform real space ansatz development. Second, our practical experience in applying CJAGP is making it increasingly clear that, in Hilbert space, the primary issue that will constrain the use of the CJAGP in the future is the fact that in its current form it must at each sample loop over a large slice of the two-electron integrals. As there has been much success in simplifying the handling of two-electron integrals in other areas of quantum chemistry, either by tensor decomposition or by screening, we look forward to the possibility of similar efficiency gains in the context of the CJAGP.

V. ACKNOWLEDGMENTS

Part of this work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. We also acknowledged support from the University of California.

Appendix A: Evaluating Eq. (19)

Here we give details on how Eq. (19) can be evaluated efficiently, assuming that p and q correspond to α spin-orbitals. The β spin case follows exactly the same logic. First, consider the case where μ_x is an orbital rotation variable, $\mu_x = K_{rs}$. For each sampled configuration, this case requires evaluation of terms of the form

$$\frac{\langle \Psi^x | a_p^\dagger a_q | \mathbf{n} \rangle}{\langle \Phi | \mathbf{n} \rangle} = \frac{\langle \Phi | (a_s^\dagger a_r - a_r^\dagger a_s) a_p^\dagger a_q | \mathbf{n} \rangle}{\langle \Phi | \mathbf{n} \rangle} \quad (\text{A1})$$

which for the different values of p, q, r, s amount to $O(N^4)$ double excitation ratios. Like those of Eq. (23), these may be evaluated for a total per-sample cost that scales as N^4 . Note that for terms with $r = p$ or $s = p$, the anti-commutation rules involved in rearranging the creation and destruction operators to match Eq. (23) will also generate single excitation ratios, but these do not change the cost scaling (see Eq. (A3) below).

It remains to consider the case where μ_x is a pairing matrix element or Jastrow coefficient. In this case it is helpful to rewrite the required term as

$$\begin{aligned} \frac{\langle \Psi^x | a_p^\dagger a_q | \mathbf{n} \rangle}{\langle \Phi | \mathbf{n} \rangle} &= \frac{\partial}{\partial \mu_x} \left(\frac{\langle \Phi | a_p^\dagger a_q | \mathbf{n} \rangle}{\langle \Phi | \mathbf{n} \rangle} \right) \\ &\quad + \frac{\langle \Phi | a_p^\dagger a_q | \mathbf{n} \rangle}{\langle \Phi | \mathbf{n} \rangle} \frac{\langle \Phi^x | \mathbf{n} \rangle}{\langle \Phi | \mathbf{n} \rangle}. \end{aligned} \quad (\text{A2})$$

From Eqs. (28-31) of Ref. [5], one can see that JAGP single excitation ratios may be evaluated as

$$\frac{\langle \Phi | a_p^\dagger a_q | \mathbf{n} \rangle}{\langle \Phi | \mathbf{n} \rangle} = (\mathbf{R}\mathbf{\Theta})_{pq} \exp(K_p^\alpha - K_q^\alpha - J_{qp}^{\alpha\alpha}), \quad (\text{A3})$$

where \mathbf{R} is the unoccupied-occupied block of the pairing matrix, $\mathbf{\Theta}$ is the inverse of the occupied-occupied block of the pairing matrix, and K_p^α and K_q^α are Ref. [5]’s Jastrow intermediates (each of which is a simple sum over Jastrow factor coefficients). As $\mathbf{\Theta}$ and the product $\mathbf{R}\mathbf{\Theta}$ are already evaluated for the JAGP LM and are thus readily available, the ratios in Eq. (A3) may all be evaluated for an additional per-sample cost scaling as N^2 . These ratios in hand, and recognizing that the pairing matrix and Jastrow derivative ratios $\langle \Phi^x | \mathbf{n} \rangle / \langle \Phi | \mathbf{n} \rangle$ are also already available, we see that the last term in Eq. (A2) may be evaluated for a per-sample cost scaling as N^4 . All that remains now is the first term on the right hand side of Eq. (A2), which requires derivatives of Eq. (A3) with respect to pairing matrix and Jastrow variables. In the Jastrow variable case, these are quite trivial, working out to $\pm \langle \Phi | a_p^\dagger a_q | \mathbf{n} \rangle / \langle \Phi | \mathbf{n} \rangle$ if the Jastrow variable appears in the exponential (remember the intermediates are just sums of Jastrow variables) and zero if it does not. For pairing matrix elements that are part of the occupied-unoccupied block \mathbf{R} for the current configura-

tion \mathbf{n} , these derivatives are

$$\frac{\partial}{\partial R_{ai}} \left(\frac{\langle \Phi | a_p^+ a_q | \mathbf{n} \rangle}{\langle \Phi | \mathbf{n} \rangle} \right) = \delta_{ap} \Theta_{iq} \exp(K_p^\alpha - K_q^\alpha - J_{qp}^{\alpha\alpha}). \quad (\text{A4})$$

For pairing matrix elements that are part of the occupied-occupied block \mathbf{O} for which Θ is the matrix inverse, these derivatives are

$$\frac{\partial}{\partial O_{ij}} \left(\frac{\langle \Phi | a_p^+ a_q | \mathbf{n} \rangle}{\langle \Phi | \mathbf{n} \rangle} \right)$$

$$= -(\mathbf{R}\Theta)_{pi} \Theta_{jq} \exp(K_p^\alpha - K_q^\alpha - J_{qp}^{\alpha\alpha}). \quad (\text{A5})$$

For other pairing matrix elements, on which the single excitation ratios do not depend, these derivatives are zero. In conclusion, whether considering orbital rotation variables via Eq. (A1) or pairing matrix or Jastrow factor variables via Eq. (A2), all of the components for a sampled configuration's contribution to \mathbf{S} via Eq. (19) may be evaluated at a cost that scales as N^4 .

-
- [1] G. K.-L. Chan and S. Sharma, *Annu. Rev. Phys. Chem.* **62**, 465 (2011).
- [2] R. E. Thomas, Q. Sun, A. Alavi, and G. H. Booth, *J. Chem. Theory Comput.* **11**, 5316 (2015).
- [3] E. Neuscamman, *J. Chem. Phys.* **139**, 181101 (2013).
- [4] R. J. Bartlett and M. Musiał, *Rev. Mod. Phys.* **79**, 291 (2007).
- [5] E. Neuscamman, *J. Chem. Phys.* **139**, 194105 (2013).
- [6] M. P. Nightingale and V. Melik-Alaverdian, *Phys. Rev. Lett.* **87**, 043401 (2001).
- [7] C. J. Umrigar, J. Toulouse, C. Filippi, S. Sorella, and R. G. Hennig, *Phys. Rev. Lett.* **98**, 110201 (2007).
- [8] J. Toulouse and C. J. Umrigar, *J. Chem. Phys.* **126**, 084102 (2007).
- [9] J. Toulouse and C. J. Umrigar, *J. Chem. Phys.* **128**, 174101 (2008).
- [10] E. Neuscamman, *Mol. Phys.*, DOI:10.1080/00268976.2015.1115903 (2015).
- [11] A. C. Hurley, J. Lennard-Jones, and J. A. Pople, *Proc. R. Soc. London, Ser. A* **220**, 446 (1953).
- [12] G. J. O. Beran, B. Austin, A. Sodt, and M. Head-Gordon, *J. Phys. Chem. A* **109**, 9183 (2005).
- [13] S. Bratož and P. Durand, *J. Chem. Phys.* **43**, 2670 (1965).
- [14] A. J. Coleman, *J. Math. Phys.* **6**, 1425 (1965).
- [15] V. N. Staroverov and G. E. Scuseria, *J. Chem. Phys.* **117**, 11107 (2002).
- [16] W. Kutzelnigg, *J. Chem. Phys.* **40**, 3640 (1964).
- [17] W. Kutzelnigg, *Theoret. chim. Acta* **3**, 241 (1965).
- [18] P. R. Surján, Ágnes Szabados, P. Jeszenszki, and T. Zoboki, *J. Math. Chem.* **50**, 534 (2012).
- [19] T. V. Voorhis and M. Head-Gordon, *J. Chem. Phys.* **112**, 5633 (2000).
- [20] T. V. Voorhis and M. Head-Gordon, *Chem. Phys. Lett.* **317**, 575 (2000).
- [21] T. V. Voorhis and M. Head-Gordon, *J. Chem. Phys.* **117**, 9190 (2002).
- [22] P. A. Limacher et al., *J. Chem. Theory Comput.* **9**, 1394 (2013).
- [23] K. Boguslawski et al., *J. Chem. Theory Comput.* **10**, 4873 (2014).
- [24] K. Boguslawski et al., *Phys. Rev. B* **89**, 201106(R) (2014).
- [25] K. Boguslawski et al., *J. Chem. Phys.* **140**, 214114 (2014).
- [26] P. Tecmer et al., *J. Phys. Chem. A* **118**, 9058 (2014).
- [27] K. Boguslawski and P. W. Ayers, *J. Chem. Theory Comput.* **11**, 5252 (2015).
- [28] V. A. Rassolov, *J. Chem. Phys.* **117**, 5978 (2002).
- [29] V. A. Rassolov, F. Xu, and S. Garashchuk, *J. Chem. Phys.* **120**, 10385 (2004).
- [30] V. A. Rassolov and F. Xu, *J. Chem. Phys.* **126**, 234112 (2007).
- [31] V. A. Rassolov and F. Xu, *J. Chem. Phys.* **127**, 044104 (2007).
- [32] B. A. Cagg and V. A. Rassolov, *J. Chem. Phys.* **141**, 164112 (2014).
- [33] P. Jeszenszki, P. R. Surján, and Ágnes Szabados, *J. Chem. Theory Comput.* **11**, 3096 (2015).
- [34] M. Casula and S. Sorella, *J. Chem. Phys.* **119**, 6500 (2003).
- [35] M. Casula, C. Attaccalite, and S. Sorella, *J. Chem. Phys.* **121**, 7110 (2004).
- [36] S. Sorella, M. Casula, and D. Rocca, *J. Chem. Phys.* **127**, 014105 (2007).
- [37] M. Marchi, S. Azadi, M. Casula, and S. Sorella, *J. Chem. Phys.* **131**, 154116 (2009).
- [38] E. G. Hohenstein, R. M. Parrish, C. D. Sherrill, and T. J. Martinez, *J. Chem. Phys.* **137**, 221101 (2012).
- [39] I. I. Ukrainskii, *Theor. Math. Phys.* **32**, 816 (1977).
- [40] J. Cullen, *Chem. Phys.* **202**, 217 (1996).
- [41] T. V. Voorhis and M. Head-Gordon, *J. Chem. Phys.* **115**, 7814 (2001).
- [42] T. Stein, T. M. Henderson, and G. E. Scuseria, *J. Chem. Phys.* **140**, 214113 (2014).
- [43] T. M. Henderson, G. E. Scuseria, J. Dukelsky, A. Signoracci, and T. Duguet, *Phys. Rev. C* **89**, 054305 (2014).
- [44] T. M. Henderson, I. W. Bulik, T. Stein, and G. E. Scuseria, *J. Chem. Phys.* **141**, 244104 (2014).
- [45] T. V. Voorhis and M. Head-Gordon, *J. Chem. Phys.* **113**, 8873 (2000).
- [46] B. Cooper and P. J. Knowles, *J. Chem. Phys.* **133**, 234102 (2010).
- [47] J. B. Robinson and P. J. Knowles, *J. Chem. Phys.* **136**, 054114 (2012).
- [48] J. B. Robinson and P. J. Knowles, *J. Chem. Theory Comput.* **8**, 2653 (2012).
- [49] J. B. Robinson and P. J. Knowles, *J. Chem. Phys.* **137**, 054301 (2012).
- [50] J. B. Robinson and P. J. Knowles, *J. Chem. Phys.* **138**, 074104 (2013).
- [51] J. Paldus and X. Li, *Adv. Chem. Phys.* **110**, 1 (1999).
- [52] T. Helgaker, P. Jørgensen, and J. Olsen, *Molecular Electronic Structure Theory*, John Wiley & Sons, Ltd., West

- Sussex, England, 2000.
- [53] Z. Bai et al., editors, *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*, SIAM, Philadelphia, 2000.
 - [54] E. R. Davidson, J. Comput. Phys. **17**, 87 (1975).
 - [55] W. E. Arnoldi, Quart. Appl. Math. **9**, 17 (1951).
 - [56] E. Neuscamman, C. J. Umrigar, and G. K.-L. Chan, Phys. Rev. B **85**, 045103 (2012).
 - [57] S. Sorella, Phys. Rev. B **64**, 024512 (2001).
 - [58] T. D. Crawford et al., J. Comput. Chem. **28**, 1610 (2007).
 - [59] H.-J. Werner and P. J. Knowles, J. Chem. Phys. **82**, 5053 (1985).
 - [60] P. J. Knowles and H.-J. Werner, Chem. Phys. Lett. **115**, 259 (1985).
 - [61] P. Knowles and N. Handy, Chem. Phys. Lett. **111**, 315 (1984).
 - [62] P. Knowles and N. Handy, Comp. Phys. Commun. **54**, 75 (1989).
 - [63] H.-J. Werner, Mol. Phys. **89**, 645 (1996).
 - [64] H.-J. Werner and P. J. Knowles, J. Chem. Phys. **89**, 5803 (1988).
 - [65] P. J. Knowles and H.-J. Werner, Chem. Phys. Lett. **145**, 514 (1988).
 - [66] H.-J. Werner et al., MOLPRO, version 2012.1, a package of ab initio programs, see <http://www.molpro.net>.
 - [67] A. Szabo and N. S. Ostlund, *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*, Dover Publications, Mineola, N.Y., 1996.
 - [68] Y. Shao et al, Phys. Chem. Chem. Phys. **8**, 3172 (2006).
 - [69] A. Krylov and P. Gill, WIREs Comput. Mol. Sci. **3**, 317 (2013).
 - [70] W. J. Hehre, R. Ditchfield, and J. A. Pople, J. Chem. Phys. **56**, 2257 (1972).
 - [71] I. W. Bulik, T. M. Henderson, and G. E. Scuseria, J. Chem. Theory Comput. **11**, 3171 (2015).
 - [72] R. Seeger and J. A. Pople, J. Chem. Phys. **66**, 3045 (1977).
 - [73] E. Neuscamman, T. Yanai, and G. K.-L. Chan, J. Chem. Phys. **130**, 124102 (2009).
 - [74] W. M. C. Foulkes, L. Mitas, R. J. Needs, and G. Rajagopal, Rev. Mod. Phys. **73**, 33 (2001).
 - [75] B. O. Roos and K. Andersson, J. Chem. Phys. **245**, 215 (1995).
 - [76] E. Miliordos and A. Mavridis, J. Phys. Chem. A **114**, 8536 (2010).
 - [77] L. Zhao and E. Neuscamman, arXiv:1508.06683 (2016).